


# When Does Differential Outcome Misclassification Matter for Estimating Prevalence?

 Jessie K. Edwards,<sup>a</sup> Stephen R. Cole,<sup>a</sup> Bonnie E. Shook-Sa,<sup>b</sup> Paul N. Zivich,<sup>a</sup> Ning Zhang,<sup>a</sup> and Catherine R. Lesko<sup>c</sup>

**Background:** When accounting for misclassification, investigators make assumptions about whether misclassification is “differential” or “nondifferential.” Most guidance on differential misclassification considers settings where outcome misclassification varies across levels of exposure, or vice versa. Here, we examine when covariate-differential misclassification must be considered when estimating overall outcome prevalence.

**Methods:** We generated datasets with outcome misclassification under five data generating mechanisms. In each, we estimated prevalence using estimators that (a) ignored misclassification, (b) assumed misclassification was nondifferential, and (c) allowed misclassification to vary across levels of a covariate. We compared bias and precision in estimated prevalence in the study sample and an external target population using different sources of validation data to account for misclassification. We illustrated use of each approach to estimate HIV prevalence using self-reported HIV status among people in East Africa cross-border areas.

**Results:** The estimator that allowed misclassification to vary across levels of the covariate produced results with little bias for both populations in all scenarios but had higher variability when the validation study contained sparse strata. Estimators that assumed nondifferential misclassification produced results with little bias when the covariate distribution in the validation data matched the covariate distribution in the target population; otherwise estimates assuming nondifferential misclassification were biased.

**Conclusions:** If validation data are a simple random sample from the target population, assuming nondifferential outcome misclassification will yield prevalence estimates with little bias regardless

of whether misclassification varies across covariates. Otherwise, obtaining valid prevalence estimates requires incorporating covariates into the estimators used to account for misclassification.

**Keywords:** Bias, Epidemiologic; Epidemiologic measurements; HIV; Validation study

(*Epidemiology* 2023;34: 192–200)

Misclassification abounds in epidemiologic studies. Numerous methods exist to account for misclassification in estimates of prevalence and associations between exposures and outcomes.<sup>1–6</sup> Such methods often require the investigator to make assumptions about whether misclassification is “nondifferential” or “differential.” When estimating associations between exposures and outcomes, the term differential is often used to imply that study design features create an expectation that the probability of misclassifying the outcome may differ by exposure status, or vice versa.<sup>7–9</sup> However, when estimating prevalence in a single sample (i.e., not stratified by exposure), misclassification probabilities may also differ across levels of other variables, some of which are predictors of the outcome.

In this work, we consider settings in which such covariate-differential misclassification is important to consider when estimating prevalence. We consider this problem in two dimensions: First, because many methods to account for misclassification rely on estimates of sensitivity and specificity from validation data, we explore scenarios where the distribution of covariates in the validation data are similar to, and different from, the main study data. Second, because we generally would like our study results to allow us to make inference to larger or different target populations, we also explore the performance of approaches to account for measurement error when transporting study results.

As a motivating example, consider estimation of HIV prevalence in the East Africa Cross-Border Integrated Health Study,<sup>10</sup> where we were interested in estimating HIV prevalence in both the study sample (thought to be representative of people socializing in cross-border areas) as well as a population composed of people socializing in cross-border areas who could be reached by a specific outreach program. In this example, the age distribution of those in the study sample was

Submitted February 17, 2022; accepted November 22, 2022

From the <sup>a</sup>Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC; <sup>b</sup>Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC; and <sup>c</sup>Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins, Baltimore, MD.

Funding: National Institutes of Health (R01AI157758, K01AA028193).

The authors report no conflicts of interest.

Code is available at <https://github.com/edwardsjk/differential>. Data used in the example are not currently available.

Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article ([www.epidem.com](http://www.epidem.com)).

Correspondence: Jessie K. Edwards, University of North Carolina at Chapel Hill, 135 Dauer Dr., 2101 McGavran-Greenberg Hall, CB #7435, Chapel Hill, NC 27599-7435. E-mail: [jessedwards@unc.edu](mailto:jessedwards@unc.edu).

ISSN: 1044-3983/23/342-192-200

DOI: 10.1097/EDE.0000000000001572

thought to differ substantially from the age distribution in the target population for the program. Self-reported HIV status is often used as a proxy for HIV serostatus<sup>11</sup> to simplify data collection, but is subject to error if patients are unaware of, or reticent to accurately report, their serostatus.<sup>12–15</sup> This reporting error may be associated with age if younger individuals were more recently infected on average and therefore had less time to receive an HIV diagnosis before the study.

In a series of simulation experiments, we explore bias and precision in estimated outcome prevalence using estimators that hold sensitivity and specificity constant across levels of covariates and using estimators that allow sensitivity and specificity to vary. We explore how performance of each estimator is affected by the source of validation data available to inform the estimators and the target population of interest.

## PART 1. EXAMPLE

### Methods

We illustrate the choices about when to allow misclassification probabilities to vary across levels of covariates when estimating HIV prevalence using self-reported HIV status from the 2016 East Africa Integrated Health Study. This study, described in detail elsewhere,<sup>10,16,17</sup> is a population-based, cross-sectional study of a wide array of health outcomes in 14 survey sites in cross-border areas in Kenya, Uganda, Tanzania, and Rwanda conducted between June 2016 and February 2017. Of the selected sites, eight were land border sites, which included the area around international border posts on highways, and six were lake border sites, which included fishing villages on Lake Victoria that served as points of commerce for fisher folk from multiple East African countries. The study included a bio-behavioral survey among a sample of people patronizing or working in public venues in cross-border areas sampled and recruited using the Priorities for Local AIDS Control Efforts method.<sup>18</sup>

The survey asked participants about past HIV testing and test results. We considered a participant to have a positive self-reported HIV status if they reported that they had received a positive HIV test result either in face-to-face interviews or in the portion of the survey where participants recorded their responses directly on study tablets. All participants (including those reporting a prior positive result) were offered a rapid HIV test at the time of the interview. Participants with a positive rapid test were linked to a local health facility for confirmatory testing and HIV care. The study was approved by ethical review boards in each country and the institutional review board at the University of North Carolina Chapel Hill.

For this analysis, we included the 9,872 participants who consented to the rapid HIV test and provided a self-reported HIV status (89% of the 11,112 survey participants). We considered the HIV rapid test as the gold standard measurement and the “true” HIV prevalence to be the proportion

of participants with a reactive HIV rapid test. Self-reported HIV status was the error-prone outcome examined. We anticipated that the sensitivity and specificity of self-reported HIV status might vary by age. To explore the importance of accounting for differential misclassification, we sampled two hypothetical validation studies from the main study sample (Table 1). Validation sample 1 was a simple random sample of size 5,000 of the main study data. Validation sample 2 also had 5,000 participants but included a greater proportion of participants over age 30 than the main study data. Such differences in the age distribution could arise if older individuals were more likely to agree to participate in the validation study or if investigators oversampled those over age 30, thought to be at higher risk of HIV, to improve precision of the estimate of overall sensitivity. Because validation studies were hypothetical (i.e., performed by sampling from the full dataset, with sampling probabilities chosen by the investigators), we knew that the probability of inclusion in validation sample 2 was affected only by age (and not other covariates).

We estimated HIV prevalence in the sample and the external target population thought to represent the population of interest for a hypothetical HIV prevention program. The population of interest for the hypothetical program was thought to differ from the study sample with respect to age only, such that 85% of people in the external target population were over 30 compared with 38% in the study sample. We applied four estimators of prevalence, described in depth below: (0) a full-data analysis, which used true HIV status for all participants (in most settings, the gold standard outcome is unobserved or partially observed<sup>19</sup>; this analysis is provided here as a comparison for the other methods); (1) a naive analysis that used self-reported HIV status as the outcome; (2) an analysis that accounted for misclassification assuming misclassification is nondifferential; and (3) an analysis that accounted for misclassification assuming misclassification varied by age. We repeated analyses 2 and 3 using both validation samples to estimate sensitivity and specificity. Except for the full-data analysis, we ignored the gold standard exposure and assumed that only the error-prone exposure was available outside of the validation sample. We used the nonparametric bootstrap in which we resampled both main study and validation sample data 500 times to construct 95% confidence intervals (CIs) around prevalence estimates.

### Estimators

Let  $Y$  represent true HIV status,  $s$  be an indicator of inclusion in the study sample ( $s = 1$ ) or the external target population ( $s = 2$ ), and  $Z$  represent age category (over 30 vs. 30 or under). In settings with no misclassification, an unbiased estimator of HIV prevalence in the study population,  $P(Y = 1|s = 1)$ , is

$$\hat{\mu}_{0,s=1} = \frac{1}{n} \sum_{i=1}^n Y_i,$$

**TABLE 1.** Counts of Participants and Proportion Reporting HIV by Age and Gold Standard HIV Status in the East Africa Cross-Border Integrated Health Study Sample and Two Hypothetical Validation Samples

| Data Source               | Age ≤ 30      |            | Age > 30      |            | Overall       |            |
|---------------------------|---------------|------------|---------------|------------|---------------|------------|
|                           | No HIV, n (%) | HIV, n (%) | No HIV, n (%) | HIV, n (%) | No HIV, n (%) | HIV, n (%) |
| Study sample <sup>a</sup> |               |            |               |            |               |            |
| No self-reported HIV      | 5,851 (99.5)  | 151 (72.2) | 3,507 (99.4)  | 140 (55.3) | 9,358 (99.4)  | 291 (63.0) |
| Self-reported HIV         | 32 (0.5)      | 58 (27.8)  | 20 (0.4)      | 113 (44.7) | 52 (0.6)      | 171 (37.0) |
| Validation dataset 1      |               |            |               |            |               |            |
| No self-reported HIV      | 2,974 (99.5)  | 79 (71.8)  | 1,775 (99.4)  | 67 (57.3)  | 4,749 (99.5)  | 146 (64.3) |
| Self-reported HIV         | 14 (0.5)      | 31 (28.2)  | 10 (0.6)      | 50 (42.7)  | 24 (0.5)      | 81 (35.7)  |
| Validation dataset 2      |               |            |               |            |               |            |
| No self-reported HIV      | 1,215 (99.4)  | 28 (66.7)  | 3,464 (99.4)  | 139 (55.2) | 4,679 (99.4)  | 167 (56.8) |
| Self-reported HIV         | 7 (0.6)       | 14 (33.3)  | 20 (0.6)      | 113 (44.8) | 27 (0.6)      | 127 (43.2) |

<sup>a</sup>Results are italicized because we would typically not observe the gold standard outcome in the entire study sample.

where the subscript  $(0, s = 1)$  indicates that this is “estimator 0” of prevalence in the study sample  $s = 1$ , and  $i = \{1, \dots, n\}$  indexes participants in the main study sample. Assuming that  $Y$  is independent of  $s$  given  $Z$ <sup>20</sup>, HIV prevalence in the external target population  $P(Y = 1|s = 2)$  may be consistently estimated by

$$\hat{\mu}_{0,s=2} = \sum_z \left\{ \hat{P}(Y = 1 | Z = z, s = 1) \hat{P}(Z = z | s = 2) \right\}.$$

In settings with misclassification, we observe some potentially mismeasured version of the outcome  $Y^*$  (here, self-reported HIV status) in place of the gold standard measurement  $Y$  (here, HIV rapid test result). It is common to ignore misclassification and estimate outcome prevalence  $P(Y = 1|s = 1)$  using the measured outcome indicator  $Y^*$  as

$$\hat{\mu}_{1,s=1} = \frac{1}{n} \sum_{i=1}^n Y_i^*$$

and  $P(Y = 1|s = 2)$  as

$$\hat{\mu}_{1,s=2} = \sum_z \left\{ \hat{P}(Y^* = 1 | Z = z, s = 1) \hat{P}(Z = z | s = 2) \right\}.$$

Here, we focus on estimators that use sensitivity  $P(Y^* = 1 | Y = 1)$  and specificity  $P(Y^* = 0 | Y = 0)$  to account for misclassification. When we have estimates of sensitivity and specificity and misclassification is not extreme (i.e., sensitivity + specificity > 1), a simple approach can be used to account for nondifferential misclassification using the observed outcome prevalence, sensitivity, and specificity,

$$\hat{\mu}_{2,s=1} = \frac{\hat{P}(Y^* = 1 | s = 1) + \hat{s}p - 1}{\hat{s}e + \hat{s}p - 1},$$

where  $\hat{P}(Y^* = 1 | s = 1) = \frac{1}{n} \sum_{i=1}^n Y_i^*$ ,  $\hat{s}e$  is the estimated sensitivity of the outcome measure, and  $\hat{s}p$  is the estimated specificity of the outcome measure.<sup>21</sup> As above, this estimator can be extended to estimate  $P(Y = 1 | s = 2)$  using

$$\hat{\mu}_{2,s=2} = \sum_z \left\{ \frac{\hat{P}(Y^* = 1 | Z = z, s = 1) + \hat{s}p_z - 1}{\hat{s}e_z + \hat{s}p_z - 1} \hat{P}(Z = z | s = 2) \right\}.$$

When sensitivity and specificity are thought to be differential across values of a covariate  $Z$ , this tabular approach can be extended:

$$\hat{\mu}_{3,s=1} = \sum_z \left\{ \frac{\hat{P}(Y^* = 1 | Z = z, s = 1) + \hat{s}p_z - 1}{\hat{s}e_z + \hat{s}p_z - 1} \hat{P}(Z = z | s = 1) \right\},$$

where  $se_z$  and  $sp_z$  represent sensitivity and specificity for individuals within stratum  $z$ . As above, this estimator can also be extended to estimate  $P(Y = 1|s = 2)$  using

$$\hat{\mu}_{3,s=2} = \sum_z \left\{ \frac{\hat{P}(Y^* = 1 | Z = z, s = 1) + \hat{s}p_z - 1}{\hat{s}e_z + \hat{s}p_z - 1} \hat{P}(Z = z | s = 2) \right\}.$$

## Example Results

Of the 9,872 participants included in the analysis, 462 tested positive for HIV on the rapid test and 223 self-reported that they had HIV. Sensitivity of self-reported HIV status as a proxy for a positive rapid test was 37% in the sample overall, and specificity was 99%. Specificity was similar across age groups, but sensitivity was 28% among those 30 or younger and 45% among those older than 30.

Using validation dataset 1 (a random draw of the main study data), we estimated that overall sensitivity was 36% (95% CI = 29, 42) and overall specificity was 99% (95% CI = 99, 100). When we stratified by age, we observed that sensitivity was lower for those 30 or younger (28%; 95% CI = 20, 37) than those over 30 (43%; 95% CI = 34, 52), and >99% specificity for both groups. Estimated sensitivity and specificity are similar to the true values reported above, though differ slightly due to sampling error.

Using validation dataset 2 (which oversampled older participants), overall sensitivity was 43% (95% CI = 38, 49) and overall specificity was 99% (95% CI = 99, 100). When we stratified by age, estimated sensitivity and specificity were similar to validation dataset 1 with 33% sensitivity among those 30 and under (95% CI = 19, 48), 45% sensitivity among those over 30 (95% CI = 39, 51), and >99% specificity for both groups.

Using the (often unavailable) gold standard measurement for all participants, estimated HIV prevalence was 4.7% (95% CI = 4.3, 5.1) in the study population and 6.2% (95% CI = 5.7, 6.7) in the external target population (Table 2). Using self-reported HIV status rather than the HIV rapid test, estimated HIV prevalence was 2.3% (95% CI = 2.0, 2.6) in the study sample and 3.2% (95% CI = 2.9, 3.6) in the external target population.

When we accounted for measurement error but assumed misclassification was nondifferential with respect to age, the validity of our results depended on the source of validation data used. We estimated that HIV prevalence in the study sample was 5.0% (95% CI = 3.6, 6.4) when using validation dataset 1, which had a similar distribution of age as the study sample, but only 4.0% (95% CI = 2.9, 5.0) when using validation dataset 2, which oversampled those over 30. Conversely, our estimate of HIV prevalence in the (older) external target population was high at 7.7% (95% CI = 5.7, 9.7) when using validation dataset 1 but 6.2% (95% CI = 4.6, 7.8) when using validation dataset 2. When we accounted for misclassification assuming it was differential by age (i.e., when we calculated sensitivity and specificity separately for those age 30 and under and for those over 30), estimated values of HIV prevalence were near the true values regardless of the source of the validation data or the target population.

## PART 2. SIMULATION EXPERIMENTS

To demonstrate when one will need to account for differential misclassification to accurately estimate prevalence, we evaluated bias and precision of estimators  $\mu_0$  to  $\mu_3$  under 4 illustrative data generating mechanisms.

## Methods

### Parameter

The parameters of interest are the prevalence of binary outcome  $Y$  in 2 distinct target populations: the population represented by the study sample  $P(Y = 1|s = 1)$  and in an external target population  $P(Y = 1|s = 2)$ .

### Misclassification Scenarios

Here, we consider four misclassification scenarios. In scenario A, observed outcome  $Y^*$  is affected only by  $Y$  and its error  $\epsilon_y$ , where  $\epsilon_y$  occurs completely at random, as shown in the diagram in Figure 1A. This would occur in our example if age affected neither true HIV prevalence nor the validity of self-reported HIV status as a proxy for true HIV status. In scenario B,  $\epsilon_y$  depends on covariate  $Z$ , but  $Z$  does not affect  $Y$  (Figure 1B). In our example, this would occur if the true HIV prevalence was not affected by age, but older people were more or less likely to correctly report their HIV status. In scenario C,  $\epsilon_y$  occurs completely at random but covariate  $Z$  affects  $Y$  (Figure 1C). In the example, this would occur if age affected true HIV status but not whether one correctly reported his or her HIV status. In scenario D,  $\epsilon_y$  is affected by  $Z$  and  $Z$  affects  $Y$  (Figure 1D). In our example, this would occur if age was associated with higher true HIV prevalence and older people were more likely to correctly report their HIV status. In scenarios B and D, misclassification varies across levels of  $Z$ . For reference, we also consider scenario 0 in which  $Z$  does not affect  $Y$ , and  $Y^*$  is a perfect proxy for  $Y$  (i.e., there is no outcome misclassification).

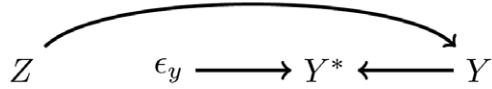
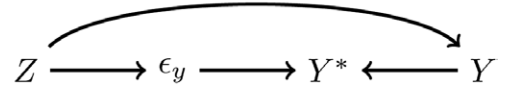
Throughout, we assume (1) we observe  $Y^*$  and  $Z$  for everyone in the study sample; (2)  $Z$  is measured without error; (3) conditional exchangeability of outcomes between the study sample and external target population, or that  $P(Y = 1 | Z, s = 1) = P(Y = 1|Z, s = 2)$ ; (4) the distribution of  $Z$  in the external target population is known; (5) neither  $Y$  nor  $Y^*$  are measured in the external target population; and (6) we have estimates of sensitivity ( $se$ ) and specificity ( $sp$ ) from a validation sample with some distribution of  $Z$  that may (or may not) differ from the study sample and from the external target population. In some simulation settings, we

**TABLE 2.** Estimated HIV Prevalence (%) Among 9,872 Participants in the East Africa Cross-Border Integrated Health Study and an External Target Population Using Three Approaches to Account for Measurement Error in Self-reported HIV Status

| Estimator  | Estimated HIV Prevalence (%) in Study Sample |                       | Estimated HIV Prevalence (%) in External Target Population |                       |
|--|--|-----------------------|--|-----------------------|
|  | Validation data 1                            | Validation data 2     | Validation data 1  | Validation data 2     |
| Full data <sup>a</sup>                                 | <i>4.7 (4.3, 5.1)</i>                        | <i>4.7 (4.3, 5.1)</i> | <i>6.2 (5.7, 6.7)</i>                                      | <i>6.2 (5.7, 6.7)</i> |
| Naive ( $\mu_1$ )                                      | 2.3 (2.0, 2.6)                               | 2.3 (2.0, 2.6)        | 3.2 (2.9, 3.6)   | 3.2 (2.9, 3.6)        |
| Assuming nondifferential misclassification ( $\mu_2$ ) | 5.0 (3.6, 6.4)                               | 4.0 (2.9, 5.0)        | 7.7 (5.7, 9.7)   | 6.2 (4.6, 7.8)        |
| Assuming differential misclassification ( $\mu_3$ )    | 4.9 (3.6, 6.3)                               | 4.3 (3.4, 6.5)        | 6.5 (4.7, 8.4)   | 6.1 (5.1, 8.0)        |

<sup>a</sup>Results are italicized because the “full data” estimator uses the gold standard outcome for all participants, which is typically unobserved, but presented here for comparison purposes.



**A****B****C****D**

**FIGURE 1.** Diagrams describing misclassification in the study sample under 4 scenarios: A) nondifferential misclassification where  $Z$  does not affect  $Y$ ; B) differential misclassification where  $Z$  does not affect  $Y$ ; C) nondifferential misclassification where  $Z$  affects  $Y$ ; and D) differential misclassification where  $Z$  affects  $Y$ .

will assume that we have  $z$ -specific estimates of sensitivity ( $se_z$ ) and specificity ( $sp_z$ ).

### Simulation Design

To understand when accounting for differential misclassification is important to accurately estimate prevalence, we

applied estimators  $\hat{\mu}_0$ ,  $\hat{\mu}_1$ ,  $\hat{\mu}_2$ , and  $\hat{\mu}_3$  of  $P(Y = 1 | s = 1)$  and  $P(Y = 1 | s = 2)$  in each of the five scenarios described above in a series of simulation experiments.

We simulated  $m = 10,000$  hypothetical study samples of size  $n = 10,000$  under each data generating mechanism in the five scenarios described above. In simulated study samples, the expected outcome prevalence was 10% in scenarios A and B and 11% in scenarios C and D, and  $Z$  was a binary covariate with expected prevalence of 75%. In simulated external target populations, the expected outcome prevalence was 10% in scenarios A and B and 16% in scenarios C and D, and the expected prevalence of  $Z$  was 33%. A supplemental set of simulations (presented in the eAppendix; <http://links.lww.com/EDE/B985>) explored a scenario with higher outcome prevalence in the main study sample.

Table 3 summarizes the relationships between  $Z$  and  $Y$  and the misclassification probabilities in each scenario. Details on data generation can be found in the Supplemental Digital Content (<http://links.lww.com/EDE/B985>). In all scenarios, we first simulated the true outcome  $Y$  and then set  $Y^*$  to be the observed (possibly misclassified) outcome. That is, if  $Y_i = 1$ , we set  $Y_i^* = 1$  with probability equal to sensitivity and if  $Y_i = 0$ , we set  $Y_i^* = 1$  with probability equal to  $1 - \text{specificity}$ .

If misclassification is differential across levels of  $Z$ , the overall sensitivity and specificity from a validation study will be a function of the prevalence of  $Z$  in the validation sample. We explored the performance of the estimators described above when informing estimates of sensitivity and specificity using eight validation studies with the same data structure

as the main study data (i.e., identical relationships between  $Z$ ,  $Y$ , and  $Y^*$ ) but different distributions of  $Z$  (Table 4), as may occur in an internal validation study with selective participation or an external validation study drawn from a population with a different distribution of  $Z$ . All validation samples had 2,000 participants. The prevalence of  $Z$  in the validation samples ranged from 15% to 85% by increments of 10%. Note validation study 7 had the same prevalence of  $Z$  as the main study sample (75%) and validation study 3 (where prevalence of  $Z$  was 35%) had similar prevalence of  $Z$  as the external target population (where prevalence of  $Z$  was 33%).

### Analysis of Simulated Data

For each simulated dataset under each scenario, we estimated the two parameters of interest,  $P(Y = 1 | s = 1)$  and  $P(Y = 1 | s = 2)$ , using the four estimators described in Part 1:

1.  $\hat{\mu}_0$ : the full-data estimator that used the typically unobserved value of  $Y$  to estimate  $P(Y = 1)$ . Because we simulated these data,  $Y$  is available for all simulated participants and the full-data estimator can be used as a reference point.

2.  $\hat{\mu}_1$ : the naive analysis that used  $Y^*$  in place of  $Y$

3.  $\hat{\mu}_2$ : estimator that accounts for so-called nondifferential misclassification

4.  $\hat{\mu}_3$ : estimator that allows sensitivity and specificity to vary by level of  $Z$

We applied estimators  $\hat{\mu}_2$  and  $\hat{\mu}_3$  using the eight types of validation studies (with different distributions of  $Z$ ) to estimate sensitivity and specificity. We report the average estimates of  $\hat{P}(Y = 1 | s = 1)$  and  $\hat{P}(Y = 1 | s = 2)$  in each scenario using each approach across the 10,000 simulated datasets, empirical bias (estimated by  $m^{-1} \sum_{j=1}^m \left\{ \hat{P}(Y = 1 | s) - \hat{P}(Y = 1 | s) \right\} \times 100$ ), empirical

**TABLE 3.** Values of Sensitivity and Specificity by Scenario in Simulated Datasets

| Scenario | Does $Z$ Affect $Y$ ? | Do Sensitivity and Specificity Vary by $Z$ ? | $Z = 1$     |             | $Z = 0$     |             |
|----------|-----------------------|--|-------------|-------------|-------------|-------------|
|          |                       |  | Sensitivity | Specificity | Sensitivity | Specificity |
| 0        | No                    | No   | 1           | 1           | 1           | 1           |
| A        | No                    | No   | 0.70        | 0.90        | 0.70        | 0.90        |
| B        | No                    | Yes  | 0.70        | 0.95        | 0.95        | 0.85        |
| C        | Yes                   | No   | 0.70        | 0.90        | 0.70        | 0.90        |
| D        | Yes                   | Yes  | 0.70        | 0.95        | 0.95        | 0.85        |

**TABLE 4.** Details of Eight Validation Studies of Size  $n = 2,000$  Explored in Simulation Experiments

| Validation Study | $P(Z=1)$ | Scenarios A and C    |                      | Scenarios B and D    |                      |
|------------------|----------|----------------------|----------------------|----------------------|----------------------|
|                  |          | Expected Sensitivity | Expected Specificity | Expected Sensitivity | Expected Specificity |
| 1                | 0.15     | 0.70                 | 0.90                 | 0.91                 | 0.87                 |
| 2                | 0.25     | 0.70                 | 0.90                 | 0.89                 | 0.88                 |
| 3                | 0.35     | 0.70                 | 0.90                 | 0.86                 | 0.89                 |
| 4                | 0.45     | 0.70                 | 0.90                 | 0.84                 | 0.90                 |
| 5                | 0.55     | 0.70                 | 0.90                 | 0.81                 | 0.91                 |
| 6                | 0.65     | 0.70                 | 0.90                 | 0.79                 | 0.92                 |
| 7                | 0.75     | 0.70                 | 0.90                 | 0.76                 | 0.93                 |
| 8                | 0.85     | 0.70                 | 0.90                 | 0.74                 | 0.94                 |

standard errors (defined as the standard deviation of the estimated prevalence across the simulated datasets), and root mean squared error (square root of bias squared plus empirical standard error squared). In addition, we provide empirical 95% confidence intervals (CIs) as the estimate  $\pm 1.96$  times the empirical standard error. We also report the maximum Monte Carlo (simulation) standard error across all scenarios as the maximum empirical standard error divided by  $\sqrt{m}$ .

## RESULTS

As expected, there was very little bias ( $<0.005$  percentage points) in estimates from any of the estimators in scenario 0, where misclassification was absent (eTable 1; <http://links.lww.com/EDE/B985>) or in any scenario using the full-data estimator (eTable 2; <http://links.lww.com/EDE/B985>).

In all scenarios with misclassification (A–D), the naive estimator was biased (Figure 2), estimating prevalence to be between 3.6 and 9.2 percentage points higher than the true value. Assuming misclassification was nondifferential (i.e., using  $\hat{\mu}_2$ ) produced results with little bias ( $<0.04$  percentage points) in scenarios A and C, where sensitivity and specificity did not vary by  $Z$ , regardless of the target population. When sensitivity and specificity varied by  $Z$  (scenarios B and D),  $\hat{\mu}_2$  still produced results with little bias if the prevalence of  $Z$  in the validation study was similar to the prevalence of  $Z$  in the target population (i.e., bias in  $\hat{P}(Y=1|s=1)$  was under 0.5 percentage points when using validation sample 7, and bias in  $\hat{P}(Y=1|s=2)$  was

under 0.6 percentage points when using validation sample 3). Estimates produced using  $\hat{\mu}_2$  were biased when sensitivity and specificity varied by  $Z$  and the distribution of  $Z$  differed between the validation study and the target population, with bias ranging from 2.35 to 7.04 percentage points. Allowing sensitivity and specificity to vary across levels of  $Z$  using  $\mu_3$  produced results with little bias ( $\leq 0.86$  percentage points) for both parameters in all scenarios, regardless of the source of validation data.

In scenarios where misclassification was nondifferential with respect to  $Z$  (scenarios A and C), estimators that allowed sensitivity and specificity to vary across levels of  $Z$  had larger empirical standard errors than estimators that assumed nondifferential misclassification. Due to increased empirical standard errors, these estimators also had higher root mean squared error than estimators that assumed nondifferential misclassification in these scenarios (Figure 3). However, in settings where misclassification was differential with respect to  $Z$ , root mean squared error was higher for estimators assuming nondifferential misclassification compared to those that allowed sensitivity and specificity to vary (due to increased bias), except for when the validation data had the same distribution of  $Z$  as the target population. Notably, when misclassification was differential, assuming nondifferential misclassification sometimes resulted in higher root mean squared error than ignoring the misclassification altogether.

Patterns in results were similar under 50% outcome prevalence in the main study sample (eTable 3; <http://links.lww.com/EDE/B985>).

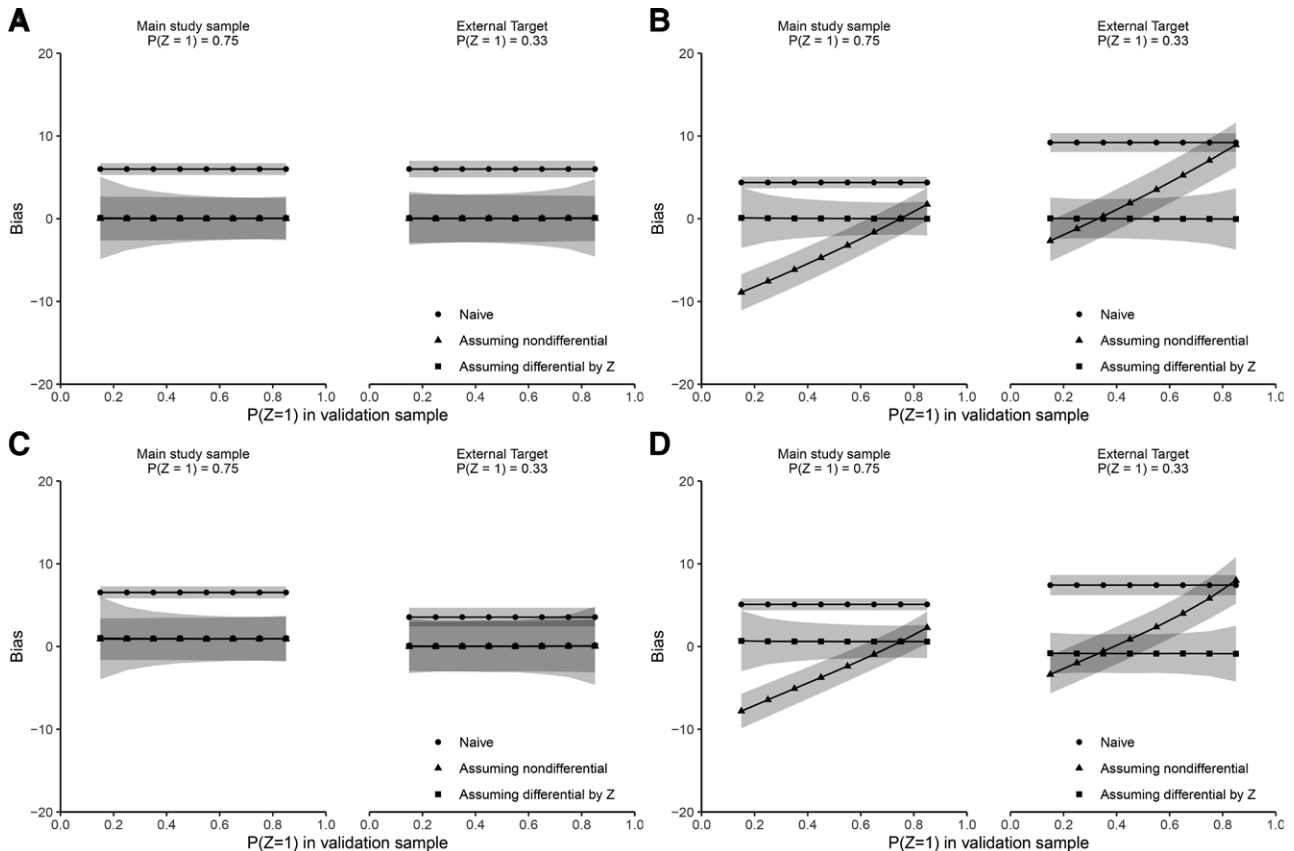
## DISCUSSION

We used data from a cross-sectional study of HIV prevalence in East Africa and a series of simulation experiments to learn about settings where accounting for covariate-differential outcome misclassification is important to estimate prevalence accurately. Our examination of assumptions about covariate-differential misclassification yielded two primary lessons. First, the source of validation data used to estimate sensitivity and specificity matters. If misclassification varies across levels of covariates, and the covariate distribution in the validation sample differs from the main study, overall estimates of sensitivity and specificity from the validation data will not be directly transportable to the main study. In this setting, estimators assuming nondifferential misclassification will yield biased results. Second, even when validation data are randomly sampled from the main study sample, prevalence estimates transported to an external target population may be biased if misclassification differs across levels of the covariates that differ between the validation sample and the external target population.

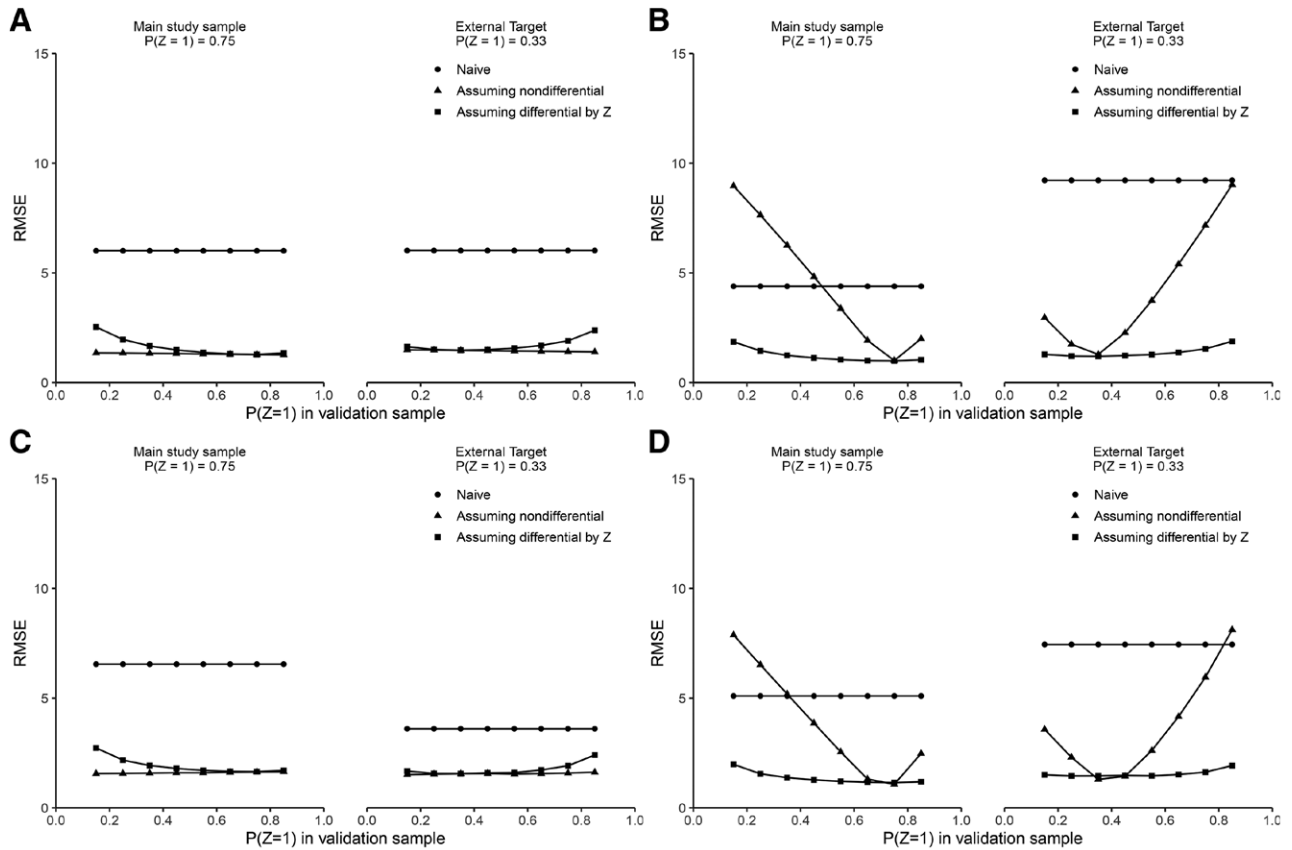
Designing validation studies or finding appropriate existing validation data is challenging. Internal validation data are rarely available, and external validation data often do not include necessary covariate data. Here, we saw that we

could share validation data across contexts without the need for covariate data in the validation sample (1) when sensitivity and specificity did not vary across levels of covariates or (2) when the distribution of covariates (across which misclassification varied) in the validation sample matched the distribution of those covariates in the target population.

A limitation of the simulations and example presented here was that we considered a single covariate  $Z$ . When sensitivity and specificity differed by  $Z$  and the distribution of  $Z$  differed between the validation sample and the target population, we saw that we could obtain an estimate with little bias by stratifying the validation data by  $Z$ , estimating sensitivity and specificity separately for each level of  $Z$ , and then stratifying the main data by  $Z$  before accounting for misclassification (estimators  $\mu_3$  described above). An alternative approach would be to reweight the validation sample to have the same distribution of  $Z$  as the target population<sup>22</sup> and then apply  $\mu_2$ , which uses single estimates of sensitivity and specificity to account for misclassification. These two approaches are about equally as onerous if  $Z$  is a single, discrete covariate. However, if  $Z$  were continuous or high dimensional, reweighting the validation study may have advantages over the stratifying by  $Z$ . For example, in the work presented here, we assumed



**FIGURE 2.** Bias in estimated prevalence under 4 misclassification scenarios in simulation studies. Bias (lines) and empirical 95% confidence intervals (shading) for estimators assuming nondifferential misclassification and misclassification differential by  $Z$  in scenarios (A–D), summarized over 10,000 simulated cohorts.



**FIGURE 3.** Weighing bias and precision in simulation studies. Root mean squared error for estimators assuming no misclassification (naive), nondifferential misclassification, and misclassification differential by  $Z$  in two target populations across scenarios (A–D), summarized over 10,000 simulated cohorts.

access to large validation studies to avoid sparse data after stratifying by  $Z$ . With small validation studies, stratifying by  $Z$  may lead to sparse data in some cells and unstable prevalence estimates. Exploration of the finite sample properties of the estimators based on validation study size was beyond our scope. However, an approach to reweight the validation data may have better accuracy in settings with smaller validation datasets because stratification by covariates would be unnecessary. Similarly, hierarchical approaches, which share some information across groups, may offer a way forward in this setting.<sup>23</sup>

Here, we applied estimators that used sensitivity and specificity to account for outcome misclassification. Other approaches account for misclassification using the positive and negative predictive values. The choice of whether to account for misclassification using sensitivity and specificity or the predictive values depends in part on the validation data or prior knowledge available to parameterize such estimators.<sup>24</sup> Examination of when predictive values must be allowed to vary across covariates is beyond the scope of this article, but we expect considerations to be similar to those examined here.

Finally, many approaches to account for misclassification, including those presented here, can be parameterized using expert knowledge or prior distributions on sensitivity and specificity in place of validation data.<sup>9,25</sup> The results presented here apply even if using these external sources of knowledge. Such knowledge is necessarily shaped by the context where it is obtained, which implies a specific covariate distribution. When parameterizing sensitivity and specificity using external knowledge, one must carefully consider the transportability of this knowledge between settings.

In conclusion, the source of information about sensitivity and specificity determines whether these parameters must be allowed to vary across covariates when accounting for outcome misclassification. If one has accurate estimates of sensitivity and specificity in the target population, estimators that assume nondifferential outcome misclassification will yield estimates of outcome prevalence in the target population with little bias. However, if estimates of sensitivity and specificity are obtained from a population with a different covariate distribution from the target population, and sensitivity and specificity differ by level of that covariate, one must account for this differential misclassification to avoid bias.



## REFERENCES

1. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *Int J Epidemiol.* 1983;12:93–97.
2. Greenland S. Variance estimation for epidemiologic effect estimates under misclassification. *Stat Med.* 1988;7:745–757.
3. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. *Epidemiology (Cambridge, Mass).* 2011;22:589–597.
4. Gravel CA, Platt RW. Weighted estimation for confounded binary outcomes subject to misclassification. *Stat Med.* 2018;37:425–436.
5. Edwards JK, Cole SR, Troester MA, Richardson DB. Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am J Epidemiol.* 2013;177:904–912.
6. Meyer MJ, Yan S, Schlageter S, Kraemer JD, Rosenberg ES, Stoto MA. Adjusting COVID-19 seroprevalence survey results to account for test sensitivity and specificity. *Am J Epidemiol.* 2022;191:681–688.
7. VanderWeele TJ, Hernán MA. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *Am J Epidemiol.* 2012;175:1303–1310.
8. Flanders WD, Drews CD, Kosinski AS. Methodology to correct for differential misclassification. *Epidemiology.* 1995;6:152–156.
9. Lash TL, Fink AK, Fox MP. Misclassification. In: Lash TL, Fox MP, Fink AK, eds. *Applying Quantitative Bias Analysis to Epidemiologic Data.* Statistics for Biology and Health. Springer; 2009:79–108.
10. Edwards JK, Arimi P, Ssengooba F, et al. The HIV care continuum among resident and non-resident populations found in venues in East Africa cross-border areas. *J Int AIDS Soc.* 2019;22:e25226.
11. Johnston LG, Sabin ML, Prybylski D, et al. The importance of assessing self-reported HIV status in bio-behavioural surveys. *Bull World Health Organ.* 2016;94:605–612.
12. Latkin CA, Vlahov D. Socially desirable response tendency as a correlate of accuracy of self-reported HIV serostatus for HIV seropositive injection drug users. *Addiction.* 1998;93:1191–1197.
13. Rohr JK, Xavier Gómez-Olivé F, Rosenberg M, et al. Performance of self-reported HIV status in determining true HIV status among older adults in rural South Africa: a validation study. *J Int AIDS Soc.* 2017;20:21691.
14. Xia Y, Milwid RM, Godin A, et al. Accuracy of self-reported HIV-testing history and awareness of HIV-positive status in four sub-Saharan African countries. *AIDS.* 2021;35:503–510.
15. Mooney AC, Campbell CK, Ratlhagana MJ, et al. Beyond social desirability bias: investigating inconsistencies in self-reported HIV testing and treatment behaviors among HIV-positive adults in North West Province, South Africa. *AIDS Behav.* 2018;22:2368–2379.
16. Mulholland GE, Markiewicz M, Arimi P, Ssengooba F, Weir S, Edwards JK. HIV prevalence and the HIV treatment cascade among female sex workers in cross-border areas in East Africa. *AIDS Behav.* 2022;26:556–568.
17. Virkud AV, Arimi P, Ssengooba F, et al. Access to HIV prevention services in East African cross-border areas: a 2016–2017 cross-sectional bio-behavioural study. *J Int AIDS Soc.* 2020;23:e25523.
18. Weir SS, Pailman C, Mahlalela X, Coetzee N, Meidany F, Boerma JT. From people to places: focusing AIDS prevention efforts where it matters most. *AIDS.* 2003;17:895–903.
19. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *Int J Epidemiol.* 2015;44:1452–1459.
20. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology.* 2017;28:553–561.
21. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol.* 1978;107:71–76.
22. Ackerman B, Siddique J, Stuart EA. Calibrating validation samples when accounting for measurement error in intervention studies. *Stat Methods Med Res.* 2021;30:1235–1248.
23. MacLehose RF, Bodnar LM, Meyer CS, Chu H, Lash TL. Hierarchical semi-Bayes methods for misclassification in perinatal epidemiology. *Epidemiology.* 2018;29:183–190.
24. Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. *Int J Epidemiol.* 2020;49:1392–1396.
25. Greenland S. Bayesian perspectives for epidemiologic research: III. Bias analysis via missing-data methods. *Int J Epidemiol.* 2009;38:1662–1673.